Joint Entropy and Conditional Entropy

Consider a vector valued variable (X,Y), The joint entropy $H(X,Y)$ is

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

$$H(X,Y) = -E\{\log p(x,Y)\}$$

The conditional entropy is defined as

$$H(Y|X) = \sum_{x\in X} p(x) H(Y|X=x)$$

$$= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

Chain rule:

$$H(X,Y) = H(X) + H(Y|X)$$

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

$$= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

$$= H(X) + H(Y|X)$$

Corollary: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$ - show!

Example

X \ Y	1	2	3	4	P(Y)
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
P(X)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$H(X) = \frac{7}{4}$ Sh bits

$H(Y) = 2$ Sh bits

$$H(X|Y) = \sum_{i=1}^{4} P(Y=i)\, H(X|Y=i)$$

$$H(X|Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x|y)$$

$$= \frac{11}{8} \text{ bits}$$

$P(X=1|Y=1) \cdot P(Y=1) = P(X=1, Y=1)$

$P(X=1|Y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$

Mutual Information

Consider two random variables X and Y with a joint probability distribution $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$.
The mutual information $I(X;Y)$ is relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

$$I(X;Y) = \sum_{x\in X} \sum_{y\in Y} p(x,y) \cdot \log \frac{p(x,y)}{p(x)p(y)}$$

We can show that $I(X;Y) = H(X) - H(X|Y)$

$$I(X;Y) = \sum_{x\in X} \sum_{y\in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x\in X} \sum_{y\in Y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x\in X} \sum_{y\in Y} p(x,y) \log p(x) + \sum_{x\in X} \sum_{y\in Y} p(x,y) \log p(x|y)$$

$$= -\sum_{x\in X} p(x) \log p(x) - \left(-\sum_{x\in X} \sum_{y\in Y} p(x,y) \log p(x|y)\right)$$

$$= H(X) - H(X|Y)$$

- The mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.
- It follows

$$I(X;Y) = H(Y) - H(Y|X)$$

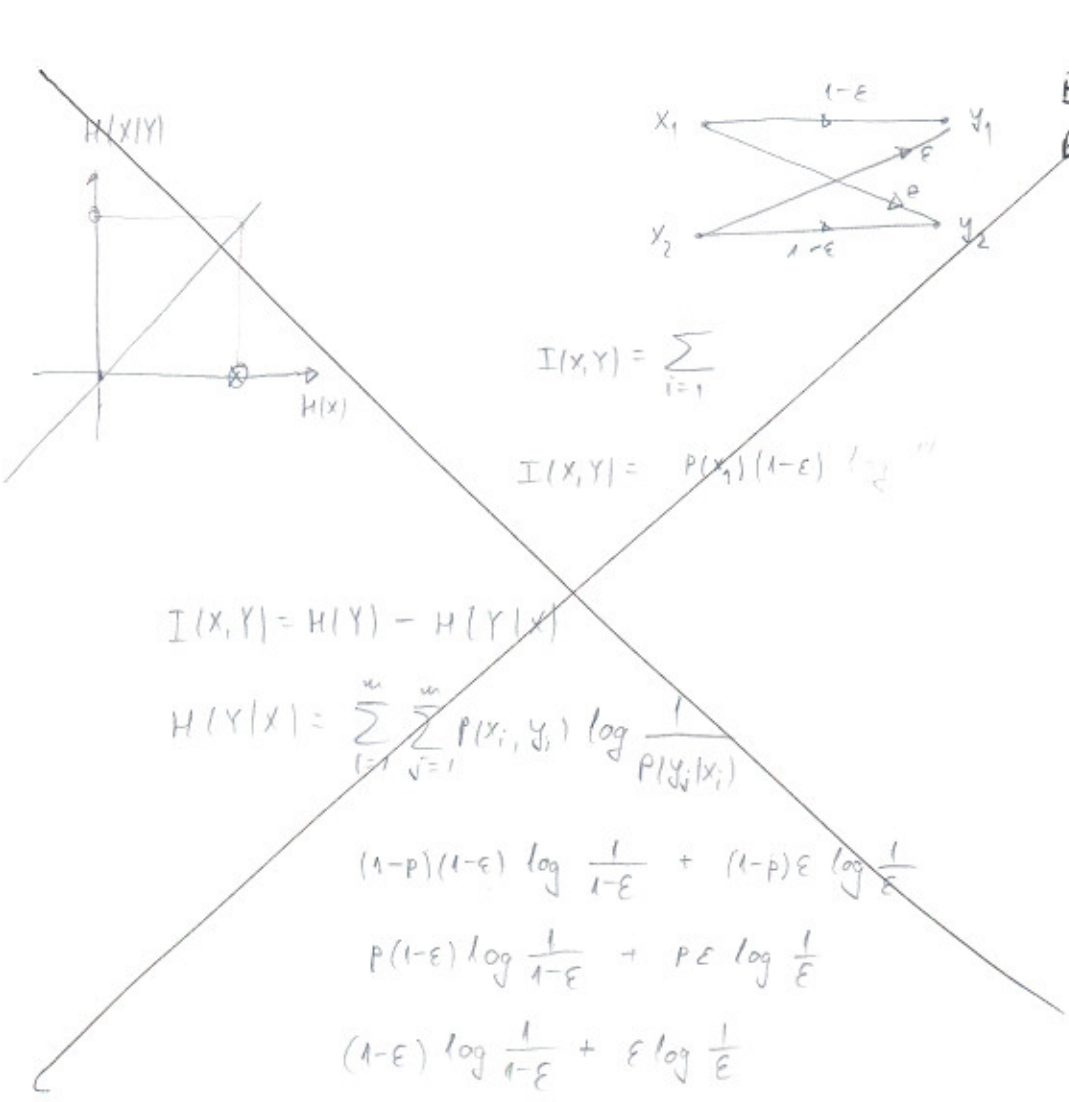
$I(X;X) = H(X)$ — $H(X)$ is referred to as self information

$$I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i) P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i) P(y_j|x_i) \log P(y_j|x_i) + \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i) P(y_j|x_i) \log P(y_j)$$

$$= \sum_{i=1}^{m} P(x_i) \sum_{j=1}^{n} P(y_j|x_i) \log P(y_j|x_i) + \sum_{j=1}^{n} \log P(y_j) \underbrace{\sum_{i=1}^{m} P(x_i) P(y_j|x_i)}_{P(y_j)}$$

$$= -H(Y|X) + H(Y)$$



$$I(X,Y) = \sum_{i=1}$$

$$I(X,Y) = P(x_1)(1-\varepsilon)$$

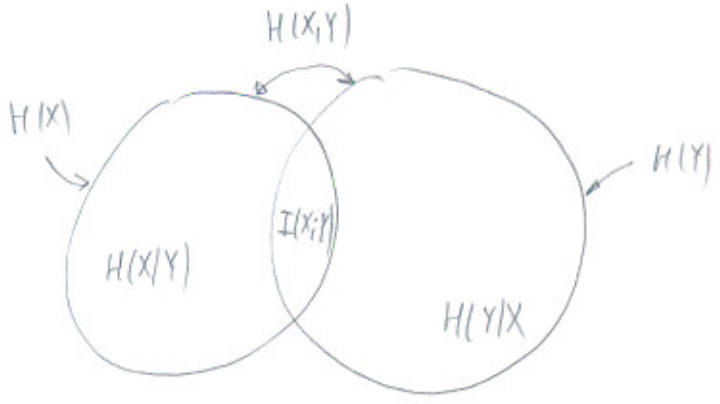
$$I(X,Y) = H(Y) - H(Y|X)$$

$$H(Y|X) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_i) \log \frac{1}{P(y_j|x_i)}$$

$$(1-p)(1-\varepsilon) \log \frac{1}{1-\varepsilon} + (1-p)\varepsilon \log \frac{1}{\varepsilon}$$

$$p(1-\varepsilon) \log \frac{1}{1-\varepsilon} + p\varepsilon \log \frac{1}{\varepsilon}$$

$$(1-\varepsilon) \log \frac{1}{1-\varepsilon} + \varepsilon \log \frac{1}{\varepsilon}$$



$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Some interesting properties:

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \dots + H(X_n|X_1, X_2, \dots X_{n-1})$$

$$I(X_1, X_2, \dots, X_n; Y) = I(X_1;Y) + I(X_2;Y|X_1) + I(X_3;Y|X_1, X_2) + \dots + (X_n;Y|X_1, X_{n-1})$$

$$I(X;Y) \geq 0$$

$$H(X|Y) \leq H(X)$$

$$H(X_1, X_2, \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Data processing inequality

If X, Y, Z form a Markov chain ($P(X,Y,Z) = P(X)\,P(Y|X)\,P(Z|Y)$) then

$$I(X;Y) \geq I(X;Z)$$

$$I(X;Y,Z) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0}$$

$$\text{or} \quad = I(X;Y) + I(X;Z|Y)$$

$$I(X;Y) \geq I(X;Z)$$

- since X and Z are conditionally independent given Y, $I(X;Z|Y) = 0$
- since $I(X;Y|Z) \geq 0$

$$I_{X;Y}(x_i;y_j) = \log \frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)}$$

$$I_{X;Y}(x_i;y_j) = \log \frac{P_{X|Y}(x_i|y_j)\, P(y_j)}{P_X(x_i)\, P(y_j)}$$

$$= \log \frac{P_{XY}(x_i,y_j)}{P_X(x_i)\, P(y_j)} = \log \frac{P_{Y|X}(y_j|x_i)\, P(x_i)}{P(x_i)\, P_Y(y_j)}$$

$$= \log \frac{P_{Y|X}(y_j|x_i)}{P(y_j)}$$

$$= I_{Y;X}(y_j;x_i)$$

$$I(X;Y) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{XY}(x_i,y_j) \log \frac{P_{XY}(x_i,y_j)}{P_X(x_i)\, P_Y(y_j)}$$

$$I(u_1;u_2|u_3) = \log \frac{P(u_1|u_2,u_3)}{P(u_1|u_3)} = I(u_1|u_3) - I(u_1|u_2,u_3)$$

$$I(U_1;U_2|U_3) = \sum_{u_1} \sum_{u_2} \sum_{u_3} P(u_1,u_2,u_3) \cdot \log \frac{P(u_1|u_2,u_3)}{P(u_1|u_3)}$$

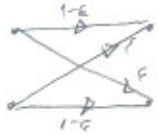
$$= H(U_1|U_3) - H(U_1|U_2U_3)$$

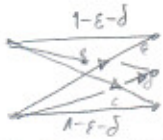
Average Mutual Information

Discrete Memoryless Channel

Channel Models

Binary Symmetric Channel





Binary Erasure channel

$P(x_i) \quad 1 \le i \le m$

$P(y_j) \quad 1 \le j \le m$

$P(y_j | x_i)$

$$H(X) = \sum_{i=1}^{m} P(x_i) \log \frac{1}{P(x_i)}$$

$$H(Y) = \sum_{j=1}^{m} P(y_j) \log \frac{1}{P(y_j)}$$

mutual information

$$I(x_i; y_j) = \log \frac{P(x_i | y_j)}{P(x_i)}$$ – information provided about x_i by y_j

$$I(x_i; y_j) = \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)}$$

amount of average / uncertainty remaining in X after observation of Y

$$H(X|Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) \log \frac{1}{P(x_i | y_j)}$$

$$I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) I(x_i; y_j)$$

← average mutual information

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(Y) - H(Y|X)$$

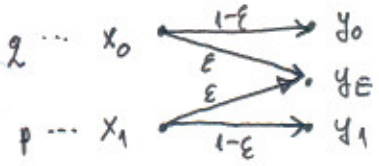
$$I(U_1; U_2 | U_3) = \sum_{u_1} \sum_{u_2} \sum_{u_3} P(u_1, u_2, u_3) \log \frac{P(u_1 | u_2, u_3)}{P(u_1 | u_3)}$$

$$I(X,Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i) P(y_j | x_i) \log \frac{P(y_j | x_i)}{P(y_j)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) \log P(y_j | x_i) - \sum_{j=1}^{m} \log P(y_j) \underbrace{\sum_{i=1}^{m} P(x_i | P(y_j | x_i)}_{P(y_j)}$$

$$= -H(Y|X) + H(Y)$$

BEC



$$P(x_0,y_0) = P(x_0)\cdot P(y_0|x_0) = q(1-\varepsilon)$$
$$P(x_0,y_E) = P(x_0)\,P(y_E|x_0) = q\cdot\varepsilon$$
$$P(x_0,y_1) = P(x_0)\,\underbrace{P(y_1|x_0)}_{0} = 0$$

$$P(x_1,y_0) = 0$$
$$P(x_1,y_E) = p\,\varepsilon$$
$$P(x_1,y_1) = P(1-\varepsilon)$$

$$P(y_0) = \sum_i P(x_i)\,P(y_0|x_i) = q\cdot(1-\varepsilon)$$
$$P(y_E) = \sum_i P(x_i)\,P(y_E|x_i) = q\varepsilon + p\varepsilon = \varepsilon$$
$$P(y_1) = \sum_i P(x_i)\,P(y_1|x_i) = p(1-\varepsilon)$$

$$I(X;Y) = H(Y) - H(Y|X)$$

$$H(Y) = -\sum_j P(y_j)\log P(y_j) = -q(1-\varepsilon)\log q(1-\varepsilon) - \varepsilon\log\varepsilon - p(1-\varepsilon)\log p(1-\varepsilon)$$

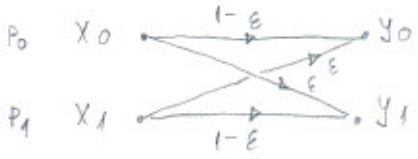
$$H(Y,X) = -\sum_i\sum_j P(x_i,y_i)\log P(y_j|x_i)$$
$$= -q(1-\varepsilon)\cdot\log(1-\varepsilon) - \varepsilon\log\varepsilon - p\cdot\varepsilon\log\varepsilon - p(1-\varepsilon)\log(1-\varepsilon)$$

$$H(Y) - H(Y,X) = -q(1-\varepsilon)\log q - q(1-\varepsilon)\log(1-\varepsilon) - p(1-\varepsilon)\log p - p(1-\varepsilon)\log(1-\varepsilon) + q(1-\varepsilon)\log(1-\varepsilon) + p(1-\varepsilon)\log(1-\varepsilon) - \varepsilon\log\varepsilon + \varepsilon\log\varepsilon$$
$$= -q(1-\varepsilon)\log q - p(1-\varepsilon)\log p$$

$$I(X;Y) = -(1-\varepsilon)(q\log q - p\log p)$$
$$= (1-\varepsilon)H(X) = H(X) - \varepsilon H(X) \qquad (H(X|Y) = \varepsilon H(X))$$

$$I_{max}(X;Y) = 1-\varepsilon$$

Binary Symmetric Channel



$$I(X;Y) = H(Y) - H(Y|X)$$

$$P(y_0) = P(x_0)\cdot(1-\varepsilon) + P(x_1)\cdot\varepsilon = P(y_0) + [P(y_0)-P(x_1)]\varepsilon$$

$$P(y_1) = P(x_1) + [P(x_1)-P(y_0)]\varepsilon$$

$$H(Y|X) = \sum_{i=0}^{1}\sum_{j=0}^{1} P(x_i,y_j)\log\frac{1}{P(y_j|x_i)} = \sum_{i=0}^{1}\sum_{j=0}^{1} P(x_i)\,P(y_j|x_i)\log\frac{1}{P(y_j|x_i)}$$

$$H(Y|X) = P_0(1-\varepsilon)\log\frac{1}{1-\varepsilon} + P_0\cdot\varepsilon\log\frac{1}{\varepsilon} + P_1\varepsilon\log\frac{1}{\varepsilon} + P_1(1-\varepsilon)\log\frac{1}{1-\varepsilon}$$

$P_0 = P_1 = \frac{1}{2}$ maximale

$H(Y|X) = (1-\varepsilon)\log\frac{1}{1-\varepsilon} + \varepsilon\log\frac{1}{\varepsilon} = H(\varepsilon)$

$H(Y) = 1$

$$I(X;Y) = 1 - H(\varepsilon)$$

How to chose input probabilities to maximize $I(X,Y)$
(DMC)

$$I(X_i;Y) = \sum_{j=1}^{n} P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)}$$

$$I(X;Y) = \sum_{i=1}^{m} P(x_i)\, I(y_j|x_i)$$

$$I(X_i;Y) = \begin{cases} = C & P(y_i) > 0 \\ \leq C & P(y_i) = 0 \end{cases}$$

Arimoto-Blahut algorithm

Example: Symmetric channels

- Each row of Π contains the same set of probabilities $\{p_j\}_{1 \le j \le n}$ (*) and each column contains the same set of numbers $\{q_i\}_{1 \le i \le m}$ (*)
- Examples:

$$\Pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{bmatrix} \qquad \Pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

- An important property of the symmetric channel is that $H(Y|X)$ is independent of $P(X_i)$ and depends only on the channel transition probability matrix.

$$H(Y|X) = -\sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i)\, P(y_j|x_i) \log \frac{1}{P(y_j|x_i)}$$

$$= -\sum_{i=1}^{m} P(x_i) \underbrace{\sum_{j=1}^{n} P(y_j|x_i) \log \frac{1}{P(y_j|x_i)}}_{\text{independent on } i} \quad \Delta$$

$$= -\underbrace{\sum_{i=1}^{m} P(x_i)}_{=1} \underbrace{\sum_{j=1}^{n} p_j \log \frac{1}{p_j}}_{\text{still independent on } i}$$

$$= -\sum_{j=1}^{n} p_j \log \frac{1}{p_j}$$

- We have proved that the conditional entropy $H(Y|X)$ does not depend on the input probability distribution. Thus the problem of maximizing $I(X;Y) = H(Y) - H(Y|X)$ reduces to the problem of maximizing the output entropy $H(Y)$.
- We know that $H(Y) \leq \log_2 n$ where the equality is achieved when $P(y_j) = \frac{1}{n} \quad 1 \leq j \leq n$
- We prove that the output symbols are equally likely when so are input symbol

$$P(y_j) = \sum_{i=1}^{m} P(x_i) P_{ij} = \sum_{i=1}^{m} P(x_i) \cdot Q_i \quad \Rightarrow \quad \text{if } P(x_i) = \frac{1}{m}$$

(A) ↑ independent on j.

$$P(y_j) = \frac{1}{m} \sum_{i=1}^{m} Q_i \text{ — all equal}$$

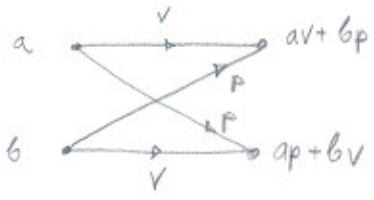
- Thus all symbols $y \in Y$ have the same probability and the capacity of a symmetric channel is given by

$$C = \log n + \sum_{j=1}^{n} P_j \log_2 P_j$$

-

Example

Binary Symetric Channel (BSC)



$$I(X;Y) = H(Y) - H(Y|X)$$

$$= H(Y) - \left(a \cdot v \log \frac{1}{v} + a \cdot p \log \frac{1}{p} + b\, p \log \frac{1}{p} + b\, v \log \frac{1}{v} \right)$$

$$= H(Y) - (a+b)\underbrace{\left(p \log \frac{1}{p} + v \log \frac{1}{v} \right)}_{H(p)}$$

$$= H(ap+bv) - H(p)$$

$$H(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$$

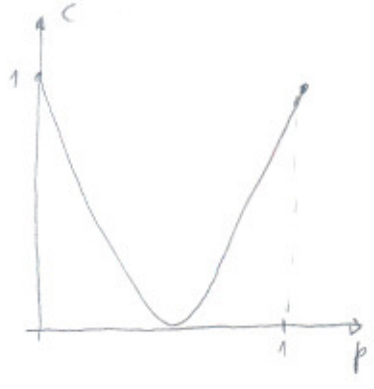
$$\left.\begin{array}{ll} p \le ap+bv \le v & \text{for } p \le \frac{1}{2} \\ p \ge ap+bv \ge v & \text{for } p \ge \frac{1}{2} \end{array}\right\} \Rightarrow H(ap+bv) \ge H(p)$$

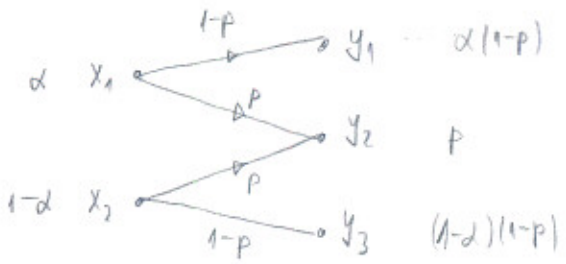
$$\Rightarrow I(X;Y) \ge 0$$

For a fixed p $H(ap+bv)$ achieves maximum for $a = \frac{1}{2}$

$$H(ap+bv) = H\left(\tfrac{1}{2}p + \tfrac{1}{2}(1-p)\right) = H\left(\tfrac{1}{2}\right) = 1 \ \frac{\text{bit}}{\text{symb}}$$

$$I_{max}(X;Y) = 1 - H(p) \qquad C = 1 - H(p)$$





$$I(X;Y) = H(Y) - H(Y|X)$$

$$= -\alpha(1-p)\log\alpha(1-p) - p\log p - (1-\alpha)(1-p)\log(1-\alpha)(1-p)$$

$$+ \alpha(1-p)\log(1-p) + \alpha p\log p + (1-\alpha)p\log p + (1-\alpha)(1-p)\log(1-p)$$

$$\Rightarrow$$

$$C = \max_{p(x_i)} I(X;Y) = I(X;Y)\Big|_{\alpha=\frac{1}{2}} = 1-p$$